Metabarcoding and Metagenomics 5: 163–174 DOI 10.3897/mbmg.5.69657

Software Description

The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data

Haris Zafeiropoulos[1,2], **Laura Gargan**[3], **Sanni Hintikka**[3], **Christina Pavloudi**[2], **Jens Carlsson**[3]

1 *Department of Biology, University of Crete, Voutes University Campus, Heraklion, Greece*

2 *Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Heraklion, Crete, Greece*

3 *Area52 Research Group, School of Biology and Environmental Science/Earth Institute, University College Dublin, Dublin 4, Ireland*

Corresponding author: Haris Zafeiropoulos (haris-zaf@hcmr.gr)

Academic editor: Alexander Probst | **Received** 9 June 2021 | **Accepted** 20 October 2021 | **Published** 03 November 2021

Abstract

The mitochondrial cytochrome C oxidase subunit I gene (COI) is commonly used in environmental DNA (eDNA) metabarcoding studies, especially for assessing metazoan diversity. Yet, a great number of COI operational taxonomic units (OTUs) or/and amplicon sequence variants (ASVs) retrieved from such studies do not get a taxonomic assignment with a reference sequence. To assess and investigate such sequences, we have developed the Dark mAtteR iNvestigator (DARN) software tool. For this purpose, a reference COI-oriented phylogenetic tree was built from 1,593 consensus sequences covering all the three domains of life. With respect to eukaryotes, consensus sequences at the family level were constructed from 183,330 sequences retrieved from the Midori reference 2 database, which represented 70% of the initial number of reference sequences. Similarly, sequences from 431 bacterial and 15 archaeal taxa at the family level (29% and 1% of the initial number of reference sequences respectively) were retrieved from the BOLD and the PFam databases. DARN makes use of this phylogenetic tree to investigate COI pre-processed sequences of amplicon samples to provide both a tabular and a graphical overview of their phylogenetic assignments. To evaluate DARN, both environmental and bulk metabarcoding samples from different aquatic environments using various primer sets were analysed. We demonstrate that a large proportion of non-target prokaryotic organisms, such as bacteria and archaea, are also amplified in eDNA samples and we suggest prokaryotic COI sequences to be included in the reference databases used for the taxonomy assignment to allow for further analyses of dark matter. DARN source code is available on GitHub at https://github.com/hariszaf/darn and as a Docker image at https://hub.docker.com/r/hariszaf/darn.

Key Words

Docker, environmental DNA (eDNA), metabarcoding, mitochondrial cytochrome C oxidase subunit I, software tool, tree of life (tol), unassigned sequences

Author summary

DARN is a software approach aiming to provide further insight into COI amplicon data of environmental samples. Building a COI-oriented reference phylogenetic tree is a challenging task especially considering the small number of microbial curated COI sequences deposited in reference databases; e.g. ~4,000 bacterial and ~150 archaeal sequences in BOLD. Inevitably, as more and more such sequences are collated, the DARN approach improves. To provide a more interactive way of communicating both our approach and our results, we strongly suggest the reader to visit this Google Collab notebook where the building of the reference COI phylogenetic tree is described step-by-step and also this GitHub pages site where our results are demonstrated. Our approach corroborates the known presence of microbial sequences in COI environmental sequencing samples and highlights the need for curated bacterial and archaeal COI sequences and their integration into reference databases (i.e., Midori, BOLD etc). We argue that DARN will benefit researchers as a quality control tool for their sequenced samples

Copyright *Haris Zafeiropoulos et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

in terms of distinguishing eukaryotic from non-eukaryotic OTUs/ASVs, but also in terms of understanding the known unknowns. As the cover ratio of COI sequences of the known taxa increases, approaches such as the one used in this study, will also enable the identification/prediction of unknown unknowns.

Introduction

Metabarcoding: concept and caveats

DNA metabarcoding is a rapidly evolving method that is being more frequently employed in a range of fields, such as biodiversity, biomonitoring, molecular ecology and others (Deiner et al. 2017; Ruppert et al. 2019). Environmental DNA (eDNA) metabarcoding, targeting DNA directly isolated from environmental samples (e.g., water, soil or sediment, (Taberlet et al. 2012a)), is considered a holistic approach (Stat et al. 2017) in terms of biodiversity assessment, providing high detection capacity. At the same time, it allows wide-scale rapid bio-assessment (Stat et al. 2017) at a relatively low cost as compared to traditional biodiversity survey methods (Ji et al. 2013).

The underlying idea of the method is to take advantage of genetic markers, i.e. marker loci, using primers anchored in conserved regions. These universal markers should have enough sequence variability to allow distinction among related taxa and be flanked by conserved regions allowing for universal or semi-universal primer design (Deagle et al. 2014). In the case of eukaryotes, the target is most commonly mitochondrial due to higher copy numbers than nuclear DNA and the potential for species level identification. Furthermore, mitochondria are nearly universally present in eukaryotic organisms, especially in case of metazoa, and can be easily sequenced and used for identification of the species composition of a sample (Taberlet et al. 2012b). However, it is essential that comprehensive public databases containing well curated, up-to-date sequences from voucher specimens are available (Schenekar et al. 2020). This way, sequences generated by universal primers can be compared with the ones in reference databases, assessing sample OTU composition. The taxonomy assignment step of the eDNA metabarcoding method and thus, the identification via DNA-barcoding, is only as good and accurate as the reference databases (Cilleros et al. 2019).

Nevertheless, there is not a truly "universal" genetic marker that is capable of being amplified for all species across different taxa (Kress et al. 2015). Different markers have been used for different taxonomic groups (Deiner et al. 2017). While bacterial and archaeal diversity is often based on the 16S rRNA gene, for eukaryotes a diverse set of loci is used from the analogous eukaryotic rRNA gene array (e.g., ITS, 18S or 28S rRNA), chloroplast genes (for plants) and mitochondrial DNA (for eukaryotes) in an attempt for species – specific resolution (Coissac et al. 2012). The mitochondrial cytochrome c oxidase subunit I (COI) marker gene has been widely used for the barcoding of the Animalia kingdom for almost two decades (Hebert et al. 2003). There are cases where COI has been the standard marker for metabarcoding, such as in the assessment of freshwater macroinvertebrates (Elbrecht and Leese 2017) even though not all taxonomic groups can be differentiated to the species level using this locus (Deiner et al. 2017); for example, in case of fish other loci are widely used such as 12S rRNA gene (hereafter referred to as 12S rRNA) (Miya et al. 2020).

The COI locus

The mitochondrial cytochrome c oxidase subunit I (also called cox1 or/and COI) is a gene fragment of ~700 bp, widely used for metazoan diversity assessment. Here we present some of the reasons that microbial eukaryotes and prokaryotes are also amplified in such studies, raising the issue of the known unknown sequences.

COI is a fundamental part of the heme aa3-type mitochondrial cytochrome c oxidase complex: the terminal electron acceptor in the respiratory chain. Even if aa3-type Cox have been found in bacteria, there are also other cytochrome c oxidase (Cox) groups, such as the cbb3-type cytochrome c oxidases (cbb3-Cox) and the cytochrome ba3 (Ekici et al. 2012; Schimo et al. 2017).

Furthermore, the presence of highly divergent nuclear mitochondrial pseudogenes (numts) has been a widely known issue on the use of COI in barcoding and metabarcoding studies, leading to overestimates of the number of taxa present in a sample (Song et al. 2008). Numts are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms (Bensasson et al. 2001).

Thus, as Mioduchowska et al. (2018) highlight, when universal primers are used targeting the COI locus, it is possible to co-amplify both non-target numts and prokaryotes (Siddall et al. 2009). This has led to multiple erroneous DNA barcoding cases and it is now not rare to encounter bacterial sequences described as metazoan in databases such as GenBank (Mioduchowska et al. 2018).

Even though there are various known issues (Deagle et al. 2014), COI is indeed considered as the "gold standard" for community DNA metabarcoding of bulk metazoan samples (Andújar et al. 2018); bulk is an environmental sample containing mainly organisms from the taxonomic group under study providing high quality and quantity of DNA (Taberlet et al. 2018). However, as highlighted in the same study, this is not the case for eDNA samples. As Stat et al. (2017) state, in the case of eDNA samples, the target region for metazoa is found in general at considerably lower concentrations compared to those from prokaryotes because most primers targeting the COI region amplify large proportions of prokaryotes at the same time (Yang et al. 2013, 2014; Collins et al. 2019). Cold-adapted marine gammaproteobacteria are an indicative example for this case as shown by Siddall et al. (2009).

Our contribution

The co-amplification of prokaryotes explained above, is a major reason for why many Operational Taxonomic Units (OTUs) and/or Amplicon Sequence Variants (ASVs) in eDNA metabarcoding studies cannot get taxonomy assignments when metazoan reference databases are used (c.f. Aylagas et al. 2016) or they are assigned to metazoan taxa but with very low confidence estimates. Despite the presence of such OTUs/ASVs to a varying degree in metabarcoding studies using the COI marker gene (Siddall et al. 2009), to the best of our knowledge, there has not been a thorough investigation of the origin for these sequences. Although unassignable sequences could be informative, there have been few attempts to further investigate this dark matter (e.g., Sinniger et al. 2016; Haenel et al. 2017).

The aim of this study was to build a framework for extracting such non-target, potentially unassigned (or assigned with low confidence) sequences from COI environmental sequence samples, hereafter referred to as "dark matter" as per Bernard et al. (2018). We argue that the vast majority of these sequences represent microbial taxa, such as bacteria and archaea.

More specifically, based on the previously described methodology by Barbera et al. (2019) (see also full stack example of the EPA-ng algorithm) for large-scale phylogenetic placements, we built a framework to estimate to what extent the OTUs/ASVs retrieved in an environmental sample represent target taxa or not. That is, to evaluate the taxonomy assignment step in a metabarcoding analysis, by checking the phylogenetic placement of dark matter sequences. Similar studies have provided great insight into other marker genes, e.g. Jamy et al. (2020).

Implementation

Building the COI tree of life

Sequences for the COI region from all the three domains of life were retrieved from curated databases. Eukaryotic sequences were retrieved from the Midori reference 2 database (version: GB239) (Machida et al. 2017). Initially, 1,315,378 sequences were retrieved corresponding to 183,330 unique species from all eukaryotic taxa. With respect to bacteria and archaea, 3,917 bacterial COI sequences were obtained from the BOLD database (Ratnasingham and Hebert 2007). Similarly, 117 sequences from archaea were obtained from BOLD. In addition, for all the PFam protein sequences related to the accession number for COX1 (PF00115), the respective DNA sequences were extracted from their corresponding genomes. This way, an additional 217 archaeal and 9,154 bacterial sequences were obtained (see Table 1). In total, sequences from 15 archaeal, 371 bacterial families and 60 taxonomic groups of higher level not assigned in the family level, were gathered. An overview of the approach that was followed is presented in Figure 1.

Table 1. Number of sequences and taxonomic species per domain of life and resources. The (#) symbols stands for "number".

Resources	bacteria		archaea	
	# of sequences	# of strains	# of sequences	# of strains
BOLD	3,917	2,267	117	117
PFam-oriented	9,154	4,532	217	115
Total unique entries	**11,421**	**6,798**	**334**	**201**

The large number of obtained sequences effectively prevents a phylogenetic tree construction encompassing their total number in terms of building a single phylogenetic tree covering all of the three domains of life (archaea, bacteria, eukaryota). Therefore, consensus representative sequences from each of the three datasets were constructed using the PhAT algorithm (Czech et al. 2019); based on the entropy of a set of sequences, PhAT groups sequences into a given target number of groups so they reflect the diversity of all the sequences in the dataset. As PhAT uses a multiple sequence alignment (MSA) as input, all the three domain-specific datasets were aligned using the MAFFT alignment software tool v7.453 (Katoh et al. 2002).

In the case of Eukaryotes, the alignment of the corresponding sequences would be impractically long because of their large number (~183K sequences). To address this challenge, a two-step procedure was followed; a sequence subset of 500 sequences (reference set) was selected and aligned and then used as a backbone for the alignment of all the remaining eukaryotic COI sequences. All sequences were considered reliable as they were retrieved from curated databases (Midori2 and BOLD). To build the reference set, a number (n) of the longest sequences from each of the various phyla were chosen, proportionally to the number (m) of sequences of each phylum (see Suppl. material 1: Table S1). The --min-tax-level parameter of the PhAT algorithm corresponded to the class level, for the case of eukaryotes and to the family level for archaea and bacteria. This parameter forced the PhAT algorithm to build at least one consensus sequence for each class and family respectively. The taxonomy level was not the same for the case of eukaryotes sequence dataset and those of bacteria and archaea, as the number of unique eukaryotic families was one order of magnitude higher. The PhAT algorithm was invoked through the gappa v0.6.1 collection of algorithms (Czech et al. 2020).

A total of 1,109 consensus sequences (70% of total consensus sequences) were built covering the eukaryotic taxa, while 463 (29%) bacterial and 21 (1%) archaeal consensus sequences were included. The per-domain, consensus sequences returned can be found under the consensus_seqs directory on the GitHub repository (see _consensus.fasta files). These sequences were then merged as a single dataset and aligned to build a reference MSA; this time MAFFT was set to return using the --globalpair algorithm and the --maxiterate parameter equal to 1,000. The MSA returned was then trimmed with the ClipKIT software package (Steenwyk et al. 2020) to keep only phylogenetically informative sites. The final MSA is available on GitHub, see trimmed_all_consensus_aligned_adjust_dir.aln.

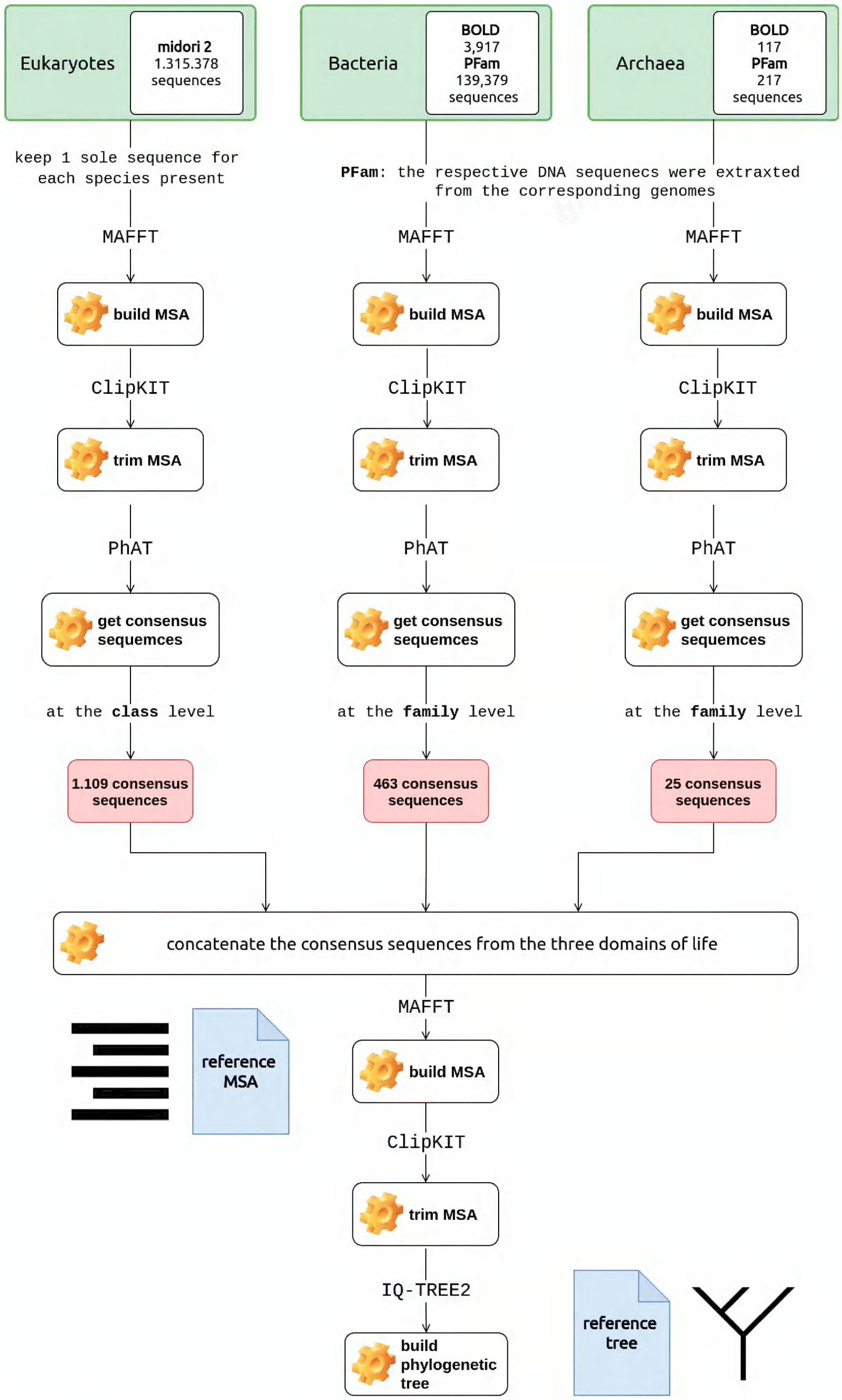

Figure 1. Overview of the approach followed to build the COI reference tree of life. Sequences were retrieved from Midori 2 (eukaryotes) and BOLD (bacteria and archaea) repositories. Consensus sequences at the family level were built for each domain specific dataset. MAFFT and consensus sequences at the family level were built using the PhAT algorithm. The COI reference tree was finally built using IQ-TREE2. Noun project icons by Arthur Slain and A. Beale.

The reference tree was then built based on this trimmed MSA using the IQ-TREE2 software (Hoang et al. 2018; Minh et al. 2020). ModelFinder was invoked through IQ-TREE2 and the GTR+F+R10 model was chosen based on the Bayesian Information Criterion (BIC) among 286 models that were tested. The phylogenetic tree was then built using 1,000 bootstrap replicates (-B 1,000) and 1,000 bootstrap replicates for Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) (-alrt 1000).

In the .iqtree file there are the branch support values; SH-aLRT support (%) / ultrafast bootstrap support (%).

A thorough description of all the implementation steps for building the reference tree is presented in this Google Collab Notebook. The computational resources of the IMBBC High Performance Computing system, called *Zorba* (Zafeiropoulos et al. 2021), were exploited to address the needs of the tasks.

Investigating COI dark matter

The COI reference tree was subsequently used to build and implement the Dark mAtteR iNvestigator (DARN) software tool. DARN uses a .fasta file with DNA sequences as input and returns an overview of sequence assignments per domain (eukaryotes, bacteria, archaea) after placing the query sequences of the sample on the branches of the reference tree. Sequences that are not assigned to a domain are grouped as "*distant*". It is necessary for the input sequences to represent the proper strand of the locus, i.e. input reads should have forward orientation. Optionally, DARN invokes the orient module of the vsearch package (Rognes et al. 2016) to implement this step, in case the user is not sure about the orientation of the sequences to be analysed.

The focal query sequences are aligned with respect to the reference MSA using the PaPaRa 2.0 algorithm (Berger and Stamatakis 2012). The query sequences are then split to build a discrete query MSA. Finally, the Evolutionary Placement Algorithm EPA-ng (Barbera et al. 2019) is used to assign the query sequences to the reference tree.

To visualise the query sequence assignments, a two-step method was developed. First, DARN invokes the gappa examine assign tool which taxonomically assigns placed query sequences by making use of the likelihood weight ratio (LWR) that was assigned to this exact taxonomic path. In the DARN framework, by making use of the --per-query-results and --best-hit flags, the gappa assign software assigns the LWR of each placement of the query sequences to a taxonomic rank that was built based on the taxonomies included in the reference tree. The first flag ensures that the gappa assign tool will return a tabular file containing one assignment profile per input query while the latter will only return the assignment with the highest LWR. DARN automatically parses this output of gappa assign to build two input Krona profile files based on a) the LWR values of each query sequence and b) an adjustive approach where all the best hits get the same value in a binary approach (presence - absence). In the final_outcome directory that DARN creates, two .html files, one for each of the Krona plots; Krona plots are built using the ktImportText command of KronaTools (Ondov et al. 2011). In addition four .fasta files are generated including the sequences of the sample that have been assigned to each domain or as "distant". A .json file with the metadata of the analysis is also returned including the identities of the sequences assigned to each domain.

DARN also runs the gappa assign tool with the --per-query-results flag only. This way, the user can have a thorough overview of each sample's sequence assignments, as a sequence may be assigned to more than one branch of the reference tree, sometimes even to different domains. However, in cases with sequences assigned to multiple branches, the likelihood scores are most typically up to 100-fold to 1000-fold different.

DARN source code as well as all data sequences and scripts for building the reference phylogenetic tree are available on GitHub.

Tree and software evaluation

Evaluation of the phylogenetic tree

The inferred phylogenetic tree is shown in Figure 2, with the bacterial (light blue) and archaeal (dark green) branches highlighted; in Suppl. material 3: Fig. S1 the distribution of the eukaryotic phyla on the tree is presented. As shown, bacteria and archaea can be distinguished from eukaryotes. Scattered bacterial branches that are present among eukaryotic ones represent the

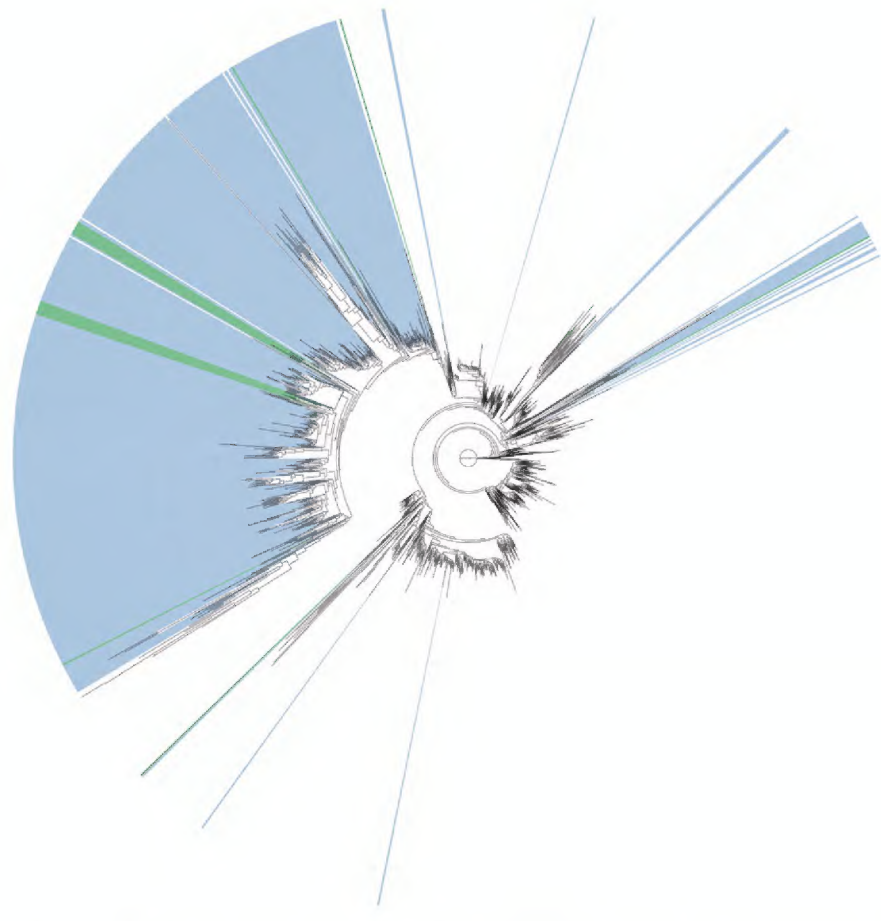

Figure 2. Phylogenetic tree of the consensus sequences retrieved; the tree that DARN makes use of. Light blue: bacterial branches. Dark green: archaeal branches. White: eukaryotic branches.

diversity of the COI locus. To evaluate the phylogenetic tree, the set of consensus sequences were placed on it using the EPA-ng algorithm. The placements (see .jplace through a phylogenetic tree viewer, e.g. iTOL) verified that the phylogenetic tree built is valid, as the consensus sequences have been placed in their corresponding taxonomic branches (Suppl. material 4: Fig. S2; the figure was built using the heat-tree module of the gappa examine tool).

DARN using mock community data

To examine whether the phylogenetic-based taxonomy assignment addresses a real-world issue, a local blast database was built using the total number of the consensus sequences retrieved. As expected, when the consensus sequences were blasted against this local blastdb, all were matched with their corresponding sequences. However, when a mock dataset was used to evaluate the two approaches (blastdb and the phylogenetic tree) none of the bacterial sequences were captured as bacteria after blastn against the local blastdb (see output file here). All bacterial sequences returned an incorrect eukaryotic assignment. Contrarily, when the phylogenetic tree was used, all the bacterial sequences were captured.

DARN using real community data

To evaluate DARN on the presence of dark matter we analysed a wide range of cases to show the ability of DARN to detect and estimate dark matter under various conditions. Both eDNA and bulk samples, from marine, lotic and lentic environments, were selected to reflect various combinations of primer and amplicon lengths, PCR protocols and bioinformatics analyses (Table 2).

More specifically, 57 marine, surface water, eDNA samples from Ireland were analysed through a. QIIME2 (Bolyen et al. 2018) and DADA2 (Callahan et al. 2016) and, b. PEMA (Zafeiropoulos et al. 2020). Similarly, 18 mangrove and 18 reef marine eDNA samples from Honduras, were analyzed using a. JAMP v0.74 (Elbrecht 2021) and DnoisE (Antich et al. 2021) and b. PEMA. Furthermore, a sediment sample and two samples from Autonomous Reef Monitoring Structures (ARMS) one conserved in DMSO and another in ethanol from the Obst et al. (2020) dataset were analysed using PEMA. In addition, one lotic and two lentic samples from Norway were analysed using PEMA. For the case of the lentic samples, multiple parameter sets regarding the ASVs inference step were implemented; i.e the *d* parameter of the Swarm v2 (Mahé et al. 2015) that PEMA invokes was set equal to 2 and 10 to cover a great

Table 2. DARN outcome over the samples or set of samples. Assignment fractions of the sequences per domain per sample in the DARN results over the samples.

Sample(s) accession number	Envo type	Sample type	Primer set	Amplicon length (bp)	Preservation method	Annealing temperature	Bioinfo pipeline(s)	# of ASVs	~ % of sequence assignments per domain (if PEMA, using *d* = 10)			
									Eukaryotes	Bacteria	Archaea	distant
ERS6449795–ERS6449829	marine	eDNA	multiplex: jgHCO2198 - jgLCO1490 and LoboF1 - LoboR1	658	water stored at 4 °C / filtered within 24 hours	60 °C × 35 cycles	QIIME2 - Dada2	13,376	11.0	**88.0**	0.02	0.003
							PEMA (d = 10)	39,454	25.0	**75.0**	0.1	0.4
ERS6463899–ERS6463901	marine reef	eDNA	mlCOIintF - jgHCO2198	313	filters stored with silica beads	46 °C × 35 cycles	JAMP dada2 PEAR vsearch DnoisE	1,304	35.0	**65.0**	**0.0**	**0.2**
ERS6463906–ERS6463911 ERS6463913–ERS6463918 ERS6463920–ERS6463922							PEMA (d = 10)	11,545	46.0	**50.0**	1.0	3.0
ERS6463744–ERS6463761	marine mangrove						JAMP dada2 PEAR vsearch DnoisE	663	40.0	**60.0**	-	**0.6**
							PEMA (d = 10)	5,879	49.0	**47.0**	**1.0**	**2.0**
ERR3460466	marine	**bulk**	mlCOIintF - jgHCO2198	313	DMSO / -20 °C	62 °C (-1 °C/ cycle) × 16 cycles and 46 °C × 24 cycles	PEMA (d = 2)	193	99.0	1.0	-	-
ERR3460467	marine	**bulk**			ETOH / -20 °C			74	97.0	**0.0**	-	3.0
ERR3460470	marine	eDNA			-20 °C			184	71.0	**28.0**	0.0	1.0
ERS6488992	lentic	eDNA	fwhF2 - EPTDr2	142	ATL-buffer	60 °C × 6 cycles and 48 °C × 26 cycles	PEMA (d = 10)	416	85.0	7.0	3.0	5.0
ERS6488993	lentic							315	99.2	0.4	0.4	-
ERS6488994	lentic							823	90.0	4.0	2.0	4.0
ERS6488995	lotic	eDNA	BF3 - BR2	458	ATL-buffer	50 °C × 35 cycles	PEMA (d = 10)	1,940	64.0	**34.0**	2.0	0.3

range of different cases (Kamenova 2020). DARN was then executed using the ASVs retrieved in each case as input. All the DARN analyses and the PEMA runs were performed on an Intel(R) Xeon(R) CPU E5649 @ 2.53GHz server of 24 CPUs and 142 GB RAM in the Area52 Research Group at the University College Dublin.

The number of sequences returned, using various bioinformatic analyses, ranged from circa 3k to 214k (Table 2) in the different amplicon datasets used. A coherent visual representation of the DARN outcome for all the datasets is available at https://hariszaf.github.io/darn/. The visual and interactive properties of the Krona plot allow the user to navigate through the taxonomy. Furthermore, DARN also supports a thorough investigation per OTU/ASV, as it returns a .json file with all the OTUs/ASVs ids that have been assigned in each of the four categories (Bacteria, Archaea, Eukaryotes and distant).

Significant proportions of non-eukaryote DARN assignments were observed in all marine eDNA samples (Table 2). Bacterial assignments made up the largest proportion of the non-eukaryotic assignments (35.3% on average and more than 75% of the OTUs/ASVs in some cases), however, archaeal assignments were also detected to a great extent as well (18.4% on average). The lentic samples were those with the shortest amplicon length among those analysed (142 bp); hence, for their orientation a database with only the shortest consensus sequences (<700 bp) was used, as otherwise a great number of sequences did not have sufficient number of hits and was discarded (see Suppl. material 2: Table S2). It is worth mentioning that in this case, the initial number of raw reads ranged from ~53,000 (ERS6488992, ERS6488993) to ~88,000 (ERS6488993) while the number of ASVs returned (using Swarm with *d* parameter equal to 10) ranged from 365 (ERS6488993) to 823 (ERS6488993). This relatively low number of ASVs could indicate that targeting such small COI regions could decrease the co-amplification of non-targeted sequences. In the case of bulk samples (Table 2) only a low proportion of the sequences were not assigned as Eukaryotes, suggesting that non-eukaryotic sequences are more abundant in environmental samples. This could be expected since prokaryotes are amplified as whole organisms from environmental samples, while metazoa that are usually the targeted taxa in COI studies, are amplified from DNA traces or/and other parts of biological source material.

Conclusions

By making use of a COI – oriented reference phylogenetic tree built from 1,593 consensus sequences, to phylogenetically place sequences from COI metabarcoding samples onto it, the surmise for including bacteria, algae, fungi etc. (Yang et al. 2013; Aylagas et al. 2016) was verified. Our results demonstrate that standard metabarcoding approaches based on the COI gene region of the mitochondrial genome will not only amplify eukaryotes, but also a large proportion of non-target prokaryotic organisms, such as bacteria and archaea. Clearly, dark matter, and especially bacteria, make up a significant proportion of sequences generated in COI based eDNA metabarcoding datasets. The large proportion of prokaryotes observed in the present study is corroborated by the findings of Yang et al. (2013). Furthermore, dark matter seems to be particularly common in eDNA as compared to bulk samples (Andújar et al. 2018). However, it should be mentioned that the high number of prokaryotic sequences in COI metabarcoding data is also reflecting known issues with contamination (Kumar et al. 2013; Dittami and Corre 2017; De Simone et al. 2020), incorrectly labeled reference sequences (Steinegger and Salzberg 2020) and holobionts (Gilbert et al. 2012; Salvucci 2016) in eukaryotic genomes.

As publicly available bacterial COI sequences are far too few to represent the bacterial and archaeal diversity, their reliable taxonomic identification is not currently possible. This way, bacterial, i.e. non-target, sequences that were amplified during the library preparation have at least the possibility of a taxonomy assignment. Our implementations using DARN indicate that it is essential both for global reference databases (e.g., BOLD, Midori etc) and custom reference databases which are commonly used, to also include non-eukaryotic sequences.

While our approach specifically addressed the COI gene, DARN can be adapted to analyse any locus fragment. For instance, metabarcoding of environmental samples for the 12S rRNA mitochondrial region is often employed to assess fish biodiversity (Weigand et al. 2019; Miya et al. 2020) and the approach presented here could be adjusted to allow further analyses of the 12S rRNA data. In addition, our approach can be used to identify non-target eukaryotes when the target is bacterial taxa (Huys et al. 2008).

The approaches implemented in DARN can benefit both bulk and eDNA metabarcoding studies, by allowing quality control and further investigation of the unassigned OTUs/ASVs. The approach is also adaptable to other markers than COI. Moreover, the approach presented here allows researchers to better understand the known unknowns and shed light on the dark matter of their metabarcoding sequence data.

Licence:

License: GNU GPLv3. For third-party components separate licenses apply.

CRediT:

H.Z. → Conceptualization, Methodology, Software, Validation, Investigation, Resources, Writing – Original draft
L.G. → Methodology, Validation, Investigation
S.H. → Validation, Investigation
C.P. → Validation, Investigation, Writing – Original draft

J.C. → Conceptualization, Investigation, Resources, Writing – Original draft

Availability

GitHub repo: https://github.com/hariszaf/darn

DockerHub repo: https://hub.docker.com/r/hariszaf/darn

The sequence data that support the findings of this study are available in the European Nucleotide Archive (ENA) with the following study accession numbers:

- Marine samples from Ireland: PRJEB45030
- Marine samples from Honduras: PRJEB45038
- Marine ARMS samples: PRJEB33796
- Lake and riverine samples from Norway: PRJEB45246

Acknowledgements

This research was supported in part through computational resources provided by IMBBC (Institute of Marine Biology, Biotechnology and Aquaculture) of the HCMR (Hellenic Centre for Marine Research). Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, LifeWatchGreece RI and the CMBR (Centre for the study and sustainable exploitation of Marine Biological Resources) RI.

In addition, sampling and sequencing of the Marine ARMS samples was funded by the infrastructure programs ASSEMBLE Plus (grant no. 730984) and the European

Marine Biological Resource Centre, EMBRC. Both programs establish and maintain the core ARMS-MBON network (http://arms-mbon.eu/) and provide services and consultation for deployment, sample processing, sequencing, data management, and analysis. The Honduran samples were partly funded by the Irish Research Council (grant no. GOIPG/2018/326), supported by Operation Wallace Ltd, UK. This research was in part funded by the Ecostructure project (part-funded by the European Regional Development Fund through the Ireland Wales Cooperation Programme 2014–2020).

We would also like to thank Dr Evangelos Pafilis (email: pafilis@hcmr.gr) for providing us access to the IMBBC HPC infrastructure, Dr Frode Fossøy (email: frode.fossoy@nina.no) for providing us environmental, lake and riverine sequence samples from Norway and the reviewers who provided thorough and fruitful comments to the manuscript.

References

Andruszkiewicz EA, Starks HA, Chavez FP, Sassoubre LM, Block BA, Boehm AB (2017) Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. PLoS ONE 12: e0176343. https://doi.org/10.1371/journal.pone.0176343

Andújar C, Arribas P, Yu DW, Vogler AP, Emerson BC (2018) Why the COI barcode should be the community DNA metabarcode for the metazoa. Molecular Ecology 27: 3968–3975. https://doi.org/10.1111/mec.14844

Antich A, Palacin C, Wangensteen OS, Turon X (2021) To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. BMC Bioinformatics 22: e177. https://doi.org/10.1186/s12859-021-04115-6

Aylagas E, Borja Á, Irigoien X, Rodríguez-Ezpeleta N (2016) Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. Frontiers in Marine Science 3: e96. https://doi.org/10.3389/fmars.2016.00096

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A (2019) EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Systematic Biology 68: 365–369. https://doi.org/10.1093/sysbio/syy054

Bensasson D, Zhang D-X, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends in Ecology & Evolution 16: 314–321. https://doi.org/10.1016/S0169-5347(01)02151-6

Berger SA, Stamatakis A (2012) PaPaRa 2.0: A Vectorized Algorithm for Probabilistic Phylogeny-Aware Alignment Extension. Heidelberg Institute for Theoretical Studies: 1–12.

Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E (2018) Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. Genome Biology and Evolution 10: 707–715. https://doi.org/10.1093/gbe/evy031

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Silva RD, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MG, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson II MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, Hooft JJ van der, Vargas F, Vázquez-Baeza Y, Vogtmann E, Hippel M von, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CH, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Inc. https://doi.org/10.7287/peerj.preprints.27295v2

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13: 581–583. https://doi.org/10.1038/nmeth.3869

Cilleros K, Valentini A, Allard L, Dejean T, Etienne R, Grenouillet G, Iribar A, Taberlet P, Vigouroux R, Brosse S (2019) Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese fresh-

water fishes. Molecular Ecology Resources 19: 27–46. https://doi.org/10.1111/1755-0998.12900

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Molecular Ecology 21: 1834–1847. https://doi.org/10.1111/j.1365-294X.2012.05550.x

Collins RA, Bakker J, Wangensteen OS, Soto AZ, Corrigan L, Sims DW, Genner MJ, Mariani S (2019) Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution 10: 1985–2001. https://doi.org/10.1111/2041-210X.13276

Czech L, Barbera P, Stamatakis A (2019) Methods for automatic reference trees and multilevel phylogenetic placement. Bioinformatics 35: 1151–1158. https://doi.org/10.1093/bioinformatics/bty767

Czech L, Barbera P, Stamatakis A (2020) Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics 36: 3263–3265. https://doi.org/10.1093/bioinformatics/btaa070

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biology Letters 10: 20140562. https://doi.org/10.1098/rsbl.2014.0562

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, Vere N de, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology 26: 5872–5895. https://doi.org/10.1111/mec.14350

De Simone G, Pasquadibisceglie A, Proietto R, Polticelli F, Aime S, JM Op den Camp H, Ascenzi P (2020) Contaminations in (meta) genome data: An open issue for the scientific community. IUBMB life 72(4): 698–705. https://doi.org/10.1002/iub.2216

Dittami SM, Corre E (2017) Detection of bacterial contaminants and hybrid sequences in the genome of the kelp *Saccharina japonica* using Taxoblast. PeerJ 5: e4073. https://doi.org/10.7717/peerj.4073

Ekici S, Pawlik G, Lohmeyer E, Koch H-G, Daldal F (2012) Biogenesis of cbb3-type cytochrome c oxidase in Rhodobacter capsulatus. Biochimica et Biophysica Acta 1817: 898–910. https://doi.org/10.1016/j.bbabio.2011.10.011

Elbrecht V (2021) JAMP: Just Another Metabarcoding Pipeline. https://github.com/VascoElbrecht/JAMP [May 28, 2021]

Elbrecht V, Leese F (2017) Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. Frontiers in Environmental Science 5: e11. https://doi.org/10.3389/fenvs.2017.00011

Gilbert SF, Sapp J, Tauber AI (2012) A symbiotic view of life: we have never been individuals. The Quarterly Review of Biology 87(4): 325–341. https://doi.org/10.1086/668166

Haenel Q, Holovachov O, Jondelius U, Sundberg P, Bourlat SJ (2017) NGS-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from Hållö island, Smögen, and soft mud from Gullmarn Fjord, Sweden. Biodiversity Data Journal 5: e12731. https://doi.org/10.3897/BDJ.5.e12731

Hebert PDN, Ratnasingham S, Waard JR de (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London – Series B: Biological Sciences 270(Suppl_1): S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35: 518–522. https://doi.org/10.1093/molbev/msx281

Huys G, Vanhoutte T, Joossens M, Mahious AS, De Brandt E, Vermeire S, Swings J (2008) Coamplification of eukaryotic DNA with 16S rRNA gene-based PCR primers: possible consequences for population fingerprinting of complex microbial communities. Current microbiology 56(6): 553–557. https://doi.org/10.1007/s00284-008-9122-z

Jamy M, Foster R, Barbera P, Czech L, Kozlov A, Stamatakis A, Bending G, Hilton S, Bass D, Burki F (2020) Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. Molecular Ecology Resources 20: 429–443. https://doi.org/10.1111/1755-0998.13117

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang X, Levi T, Lott M, Emerson BC, Yu DW (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16: 1245–1257. https://doi.org/10.1111/ele.12162

Kamenova S (2020) A flexible pipeline combining clustering and correction tools for prokaryotic and eukaryotic metabarcoding. Peer Community in Ecology 1: 100043. https://doi.org/10.24072/pci.ecology.100043

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059–3066. https://doi.org/10.1093/nar/gkf436

Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Frontiers in genetics 4: e237. https://doi.org/10.3389/fgene.2013.00237

Machida RJ, Leray M, Ho S-L, Knowlton N (2017) Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Scientific Data 4: 1–7. https://doi.org/10.1038/sdata.2017.27

Mahé F, Rognes T, Quince C, Vargas C de, Dunthorn M (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3: e1420. https://doi.org/10.7717/peerj.1420

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37: 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mioduchowska M, Czyż MJ, Gołdyn B, Kur J, Sell J (2018) Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? PLoS ONE 13: e0199609. https://doi.org/10.1371/journal.pone.0199609

Miya M, Gotoh RO, Sado T (2020) MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. Fisheries Science 86: 939–970. https://doi.org/10.1007/s12562-020-01461-x

Obst M, Exter K, Allcock AL, Arvanitidis C, Axberg A, Bustamante M, Cancio I, Carreira-Flores D, Chatzinikolaou E, Chatzigeorgiou G, Chrismas N, Clark MS, Comtet T, Dailianis T, Davies N, Deneudt K, de Cerio OD, Fortič A, Gerovasileiou V, Hablützel PI, Keklikoglou K, Kotoulas G, Lasota R, Leite BR, Loisel S, Lévêque L, Levy L, Malachowicz M, Mavrič B, Meyer C, Mortelmans J, Norkko J, Pade N, Power AM, Ramšak A, Reiss H, Solbakken J, Staehr PA, Sundberg P, Thyrring J, Troncoso JS, Viard F, Wenne R, Yperifanou

EI, Zbawicka M, Pavloudi C (2020) A Marine Biodiversity Observation Network for Genetic Monitoring of Hard-Bottom Communities (ARMS-MBON). Frontiers in Marine Science 7: 572680. https://doi.org/10.3389/fmars.2020.572680

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12: e385. https://doi.org/10.1186/1471-2105-12-385

Ratnasingham S, Hebert PDN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7: 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4: e2584. https://doi.org/10.7717/peerj.2584

Ruppert KM, Kline RJ, Rahman MS (2019) Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Global Ecology and Conservation 17: e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Salvucci E (2016) Microbiome, holobiont and the net of life. Critical Reviews in Microbiology 42(3): 485–494. https://doi.org/10.3109/1040841X.2014.962478

Schenekar T, Schletterer M, Lecaudey LA, Weiss SJ (2020) Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. River Research and Applications 36: 1004–1013. https://doi.org/10.1002/rra.3610

Schimo S, Wittig I, Pos KM, Ludwig B (2017) Cytochrome c Oxidase Biogenesis and Metallochaperone Interactions: Steps in the Assembly Pathway of a Bacterial Complex. PLoS ONE 12: e0170037. https://doi.org/10.1371/journal.pone.0170037

Siddall ME, Fontanella FM, Watson SC, Kvist S, Erséus C (2009) Barcoding Bamboozled by Bacteria: Convergence to Metazoan Mitochondrial Primer Targets by Marine Microbes. Systematic Biology 58: 445–451. https://doi.org/10.1093/sysbio/syp033

Sinniger F, Pawlowski J, Harii S, Gooday AJ, Yamamoto H, Chevaldonné P, Cedhagen T, Carvalho G, Creer S (2016) Worldwide Analysis of Sedimentary DNA Reveals Major Gaps in Taxonomic Knowledge of Deep-Sea Benthos. Frontiers in Marine Science 3: e92. https://doi.org/10.3389/fmars.2016.00092

Song H, Buhay JE, Whiting MF, Crandall KA (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences 105: 13486–13491. https://doi.org/10.1073/pnas.0803076105

Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M (2017) Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. Scientific Reports 7: e12240. https://doi.org/10.1038/s41598-017-12501-5

Steenwyk JL, Iii TJB, Li Y, Shen X-X, Rokas A (2020) ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLOS Biology 18: e3001007. https://doi.org/10.1371/journal.pbio.3001007

Steinegger M, Salzberg SL (2020) Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biology 21: e115. https://doi.org/10.1186/s13059-020-02023-1

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012a) Environmental DNA. Molecular Ecology 21: 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet P, Bonin A, Zinger L, Coissac E (2018) Analysis of bulk samples. In: Taberlet P, Bonin A, Zinger L, Coissac E (Eds) Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198767220.003.0018

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012b) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21: 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K-DB, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher JN, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of The Total Environment 678: 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Yang C, Ji Y, Wang X, Yang C, Yu DW (2013) Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity. Science China Life Sciences 56: 73–81. https://doi.org/10.1007/s11427-012-4423-7

Yang C, Wang X, Miller JA, de Blécourt M, Ji Y, Yang C, Harrison RD, Yu DW (2014) Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. Ecological Indicators 46: 379–389. https://doi.org/10.1016/j.ecolind.2014.06.028

Zafeiropoulos H, Viet HQ, Vasileiadou K, Potirakis A, Arvanitidis C, Topalis P, Pavloudi C, Pafilis E (2020) PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. GigaScience 9(3): giaa022. https://doi.org/10.1093/gigascience/giaa022

Zafeiropoulos H, Gioti A, Ninidakis S, Potirakis A, Paragkamian S, Angelova N, Antoniou A, Danis T, Kaitetzidou E, Kasapidis P, Kristoffersen JB, Papadogiannis V, Pavloudi C, Ha QV, Lagnel J, Pattakos N, Perantinos G, Sidirokastritis D, Vavilis P, Kotoulas G, Manousaki T, Sarropoulou E, Tsigenopoulos CS, Arvanitidis C, Magoulas A, Pafilis E (2021) 0s and 1s in marine molecular research: a regional HPC perspective. GigaScience 10(8): giab053. https://doi.org/10.1093/gigascience/giab053

Supplementary material 1
Table S1
Author: Haris Zafeiropoulos, Laura Gargan, Sanni Hintikka, Christina Pavloudi, Jens Carlsson
Data type: excel table
Explanation note: Number of sequences per phylum in the Eukaryotes domain dataset and the corresponding number of the longest sequences used in the 500 sequences subset (reference set) used as a backbone for the complete alignment.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.5.69657.suppl1

Supplementary material 2
Table S2
Author: Haris Zafeiropoulos, Laura Gargan, Sanni Hintikka, Christina Pavloudi, Jens Carlsson
Data type: excel table
Explanation note: Number of sequences in each DARN experiment before and after the sequence orientation step.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.5.69657.suppl2

Supplementary material 3
Figure S1
Author: Haris Zafeiropoulos, Laura Gargan, Sanni Hintikka, Christina Pavloudi, Jens Carlsson
Data type: png. image
Explanation note: Hylogenetic tree of the consensus sequences retrieved showing the distribution of the eukaryotic phyla.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.5.69657.suppl3

Supplementary material 4
Figure S2
Author: Haris Zafeiropoulos, Laura Gargan, Sanni Hintikka, Christina Pavloudi, Jens Carlsson
Data type: png. image
Explanation note: Placements of the consensus sequences used to build the COI reference phylogenetic tree for the DARN tool, onto the phylogenetic tree (stroke width for the branches of the tree is 5). The color coding represents the placements per branch, with a range from zero (blue) to a maximum of 2 (blue). The 1 leaf – 1 placement relationship, as well as the maximum of 2 placements in the color coding bar, indicate the proper placement of each consensus sequence to its corresponding branch.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.5.69657.suppl4